Technical Report No. 5/08, December 2008

ON THE SELECTION OF THE SMOOTHING PARAMETER IN POISSON SMOOTHING OF HISTOGRAM ESTIMATOR: COMPUTATIONAL ASPECTS

Yogendra P. Chaubey and Pranab K. Sen

On the Selection of the Smoothing Parameter in Poisson Smoothing of Histogram Estimator: Computational Aspects

Yogendra P. Chaubey[a1] **and Pranab K. Sen**[b]

[a]Department of Mathematics and Statistics, Concordia University, Montreal, QC, H4B 1R6, CANADA
[b]Department of Statistics and Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA

---

1 Introduction and Background

Let $\{X_i, i \geq 1\}$ be a sequence of independent and identically distributed random variables with common distribution function $F(x)$ and the density $f(x)$ supported on $\mathbb{R}^+$. Then a smooth estimator of the density function $f(x)$ is given by

$$\tilde{f}_n(x; \lambda_n, \mathcal{D}) = \lambda_n \sum_{j=0}^{N} p_j(\lambda_n x) w_j(\lambda_n, \mathcal{D}), \tag{1.1}$$

where λ_n is a constant which controls the smoothness of the estimator, $w_j(.,.), j = 1, ...$ denote the weights depending on the data $\mathcal{D}$ and the constant λ_n (see Gowronski and Stadm"uler (1980, 1981)); specifically $N = \lambda_n \max(X_1, ..., X_n)$,
$w_j(lambda, D) = F_n((j+1)/\lambda) - F_n(j/\lambda)$, F_n is the empirical distribution function based on the data $\mathcal{D}$. These estimators were independently proposed by Chaubey and Sen (1996), though the weights were truncated. The important property of the sequence $\{\lambda_n\}_{n=1}^{\infty}$ is to be chosen such that $\lambda_n \to \infty$ and $n^{-1}\lambda_n \to 0$ as $n \to \infty$.

A convenient *stochastic choice* of λ_n was proposed by Chaubey and Sen (1996) as:

$$\lambda_{n(1)} = \frac{n}{\max(X_1, ..., X_n),}$$

[1]Corresponding author: E-mail: chaubey@alcor.concordia.ca

as it satisfies the desired properties mentioned before if $E(X) < \infty$. Chaubey and Sen (1998) noticed that for the compact support this choice will not satisfy the property that $n^{-1}\lambda_n \to 0$ as $n \to infty$. To cover these cases also they proposed the choice of $\lambda_{n(2)} = n(\log\log r_n)^{-1}/X_{n-r_n+1:n}$ where $r_n = o(\log\log n)$. This choice satisfies the properties as mentioned before. We also find in the above papers that a deterministic choice

$$\lambda_{n(3)} = [n^{2/5}] + 1$$

may be recommended as in Gowronski and Stadtmüller (1981) due to the strong convergence of the density under this condition. These recommendations are based on the asymptotic theory, however, in finite sample case they may not be very satisfactory. The choice $\lambda_{n(1)}$ and $\lambda_{n(2)}$ may turn out to be very large so that they create problems in computation. The third choice may be a good guide for a first choice, but one would like to assert some 'optimality' to this constant. It is purpose of this note to investigate some cross validation methods for data adaptive choice of λ_n. Note that it is explicit in Gawronski and Stadtmüller (1981) that λ_n is an integer, but this is not necessary. We will investigate two choices of cross validation methods, one is based on the *likelihood* and the other is based on *mean integrated squared error.* The latter method is a popular one in the literature on kernel smoothing and one could expect this to be a preferred method here also. However, we have found through extensive simulations that likelihood based cross validation is numerically more convenient. Section 2 describes these methods in detail and the next section presents the results of extensive simulations. Comments on computational aspects are also detailed there. Section 4 gives conclusions of the study.

2 Likelihood and Integrated Squared Error Cross Validation

2.1 Likelihood Based Cross Validation

Bowman (1981) shows that minimizing the Kullback-Liebler divergence is equivalent to the minimization of

$$CV_{KL}(\lambda_n) = -\log\prod_{i=1}^{n}\tilde{f}_n(x;\mathcal{D}_i) = -\sum_{i=1}^{n}\log(\tilde{f}_n(x;\mathcal{D}_i)),$$

where $\mathcal{D}_i$ denotes data with X_i removed from $\mathcal{D}$. The solution of the above minimization problem will be denoted by λ_{nKL}. When the whole sample is used in constructing the discrete density estimator, we will simply denote the density by $\tilde{f}_n(x)$.

2.2 Integrated Square Error Cross Validation

According to this criterion we determine λ_n that minimizes the criterion related to the *integrated squared error*,

$$CV_{ISE}(\lambda_n) = \int \tilde{f}_n^2(x; \lambda_n^2, (D))dx - 2\frac{1}{n}\sum_{i=1}^{n} \tilde{f}_{n-1}(X_i; \lambda_n, \mathcal{D}_i).$$

The first term can be explicitly obtained and the result is shown below:

$$\int_0^\infty \tilde{f}_n(x)^2 = \frac{\lambda_n}{2}\sum_{j=0}^{N}\sum_{k=0}^{N}\frac{(j+k)!}{j!k!}\left(\frac{1}{2}\right)^{j+k} w_j(\lambda_n, \mathcal{D})w_k(\lambda_n, \mathcal{D}).$$

The solution to the above minimization problem is denoted by λ_{nISE}.

2.3 Hellinger Distance

The Hellinger distance between two densities f and g is given by

$$H(f, g) = \int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx.$$

This measure is appropriate as a benchmark to establish the closeness of the estimated density to the true density in finite samples, since this is a bounded measure unlike the measures in the above subsections. One notices that since $\int \sqrt{f(x)g(x)}dx \leq (1/2)(\int f(x)dx + \int g(x)dx) = 1/2$,

$$0 \leq H(f, g) \leq 2.$$

A value closer to 0 signifies a closer resemblance of f and g. We will use $H(\tilde{f}_n, f)$ to compare the values of different choices of λ_n for different simulated samples from f in the next section. The optimum value here will be denoted by λ_{nH}. We conjecture that as $n \to \infty$ different choices of λ_n are equivalent.

3 Simulation Studies

3.1 Some Comments on Computations

Here we simulate samples from the standard exponential distribution and Lognormal distribution for sample sizes $n = 10, 20, 30,$
$40, 50, 100, 1000$. For each sample we obtain the optimum choice of λ_n by KL and ISE cross-validation methods. To judge the closeness of the estimated density with the

true density we have listed the Hellinger distances $H(\tilde{f}_n, f)$ for each choice of λ_n. The optimum solution may be obtained using any optimizing subroutines, however, care must be taken because the function $CV(\lambda_n)$, in general is not a very smooth function due to the discrete nature of the weights $w_j(\lambda, \mathcal{D})$.

We have found that the `R-` subroutine `nlm` may not provide the correct solution, specially if the starting point is far away from the global solution. For example, Figure 1 below gives histogram of a sample from Lognormal(0,1) distribution with $max(x) = 9.102234$. The `nlm` subroutine gives 3.287119 and 6.554863 for the minima for KL and ISE cross validation respectively. The code returned is 3 for the KL criterion and 2 for the ISE criteria. These codes imply that the result is "probably the solution" and therefore we must confirm these values. The program is based on a gradient method and when the gradient tolerance is reached (or other criteria is reached) the solution is reported. Therefore, many times local minima or reported.

The default values were used, however, changing the gradient tolerance to lower label 1e-16 did not produce different results. To confirm the solution, therefore we decide to plot the criterion function in the range $[1, 20]$. These plots suggest that the correct minima for the first criterion is in the interval $[2, 5]$ and for the second criterion it is inside the interval $[2, 8]$. Hence, the reported values from the routine `nlm` seem to be correct solution. So we give the starting value 2, instead of 1. Now the reported solution is 2.002482 in both the cases. This looks reasonable in the first case but not in the second case.

Therefore, we must examine the function $CV(\lambda)$ in the neighborhood of the solution. So we decide to use the routine `optimise` which allows to input an interval for the solution. An interval of (2,5) gives the solution 2.70999 for the KL criterion and that in the interval (2,8) gives a solution of 4.294595. The solution in the first case is not for from that produced by the `nlm` routine but that in the second case is quite different. Giving a wider interval of (1,20), the solutions are respectively given by 2.629449 and 5.215846 and seem reasonable by looking at the graph. This procedure picks up the minima as 13.94625 $H(\tilde{f}_n, f)$ where as `nlm` routine gives a local minima of 6.551021 for a starting value of 1. The value of $n/max(x_1, x_2, ..., x_n)$ for this sample is given by 8.191776. The Hellinger distances of the estimated densities using Chaubey-Sen, KL, ISE with the true lognormal density are respectively given by 0.02909614, 0.03221081 and 0.02638340 which are close to true distance of 0.02316986 if we new the true density.

It may be concluded that as long as the value of λ_n is in the close neighborhood of the minima, the estimated density is not very different from the optimum choice. Corresponding four densities are plotted in Figure 2 and there is almost no difference in them qualitatively. In this particular data the plot obtained using the ISE criteria may be preferred over the others as it comes closer to the one obtained under the true minimum Hellinger distance but it is not as rough.

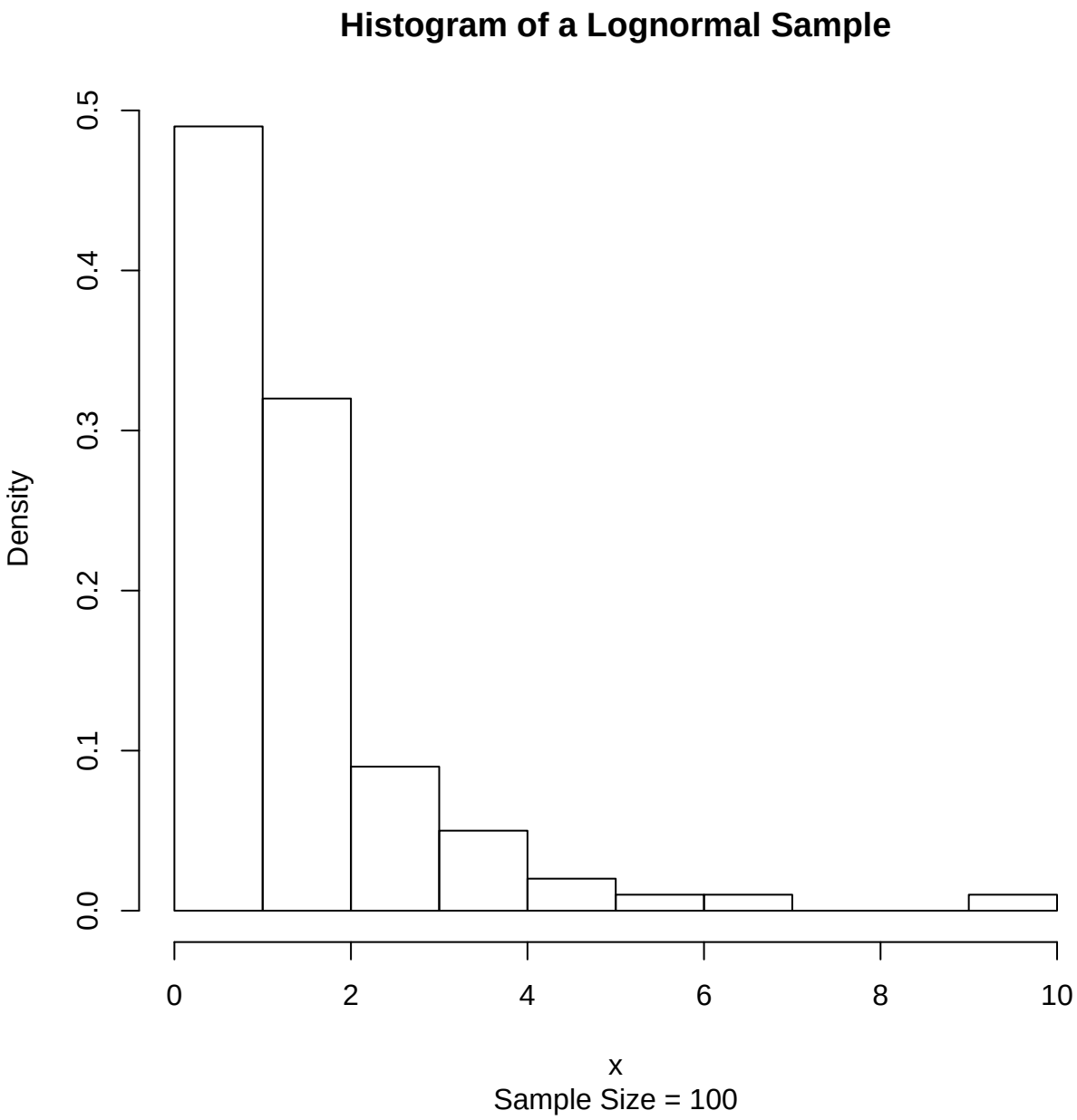

Figure 1: Histogram of a Lognormal Sample, Sample Size = 100

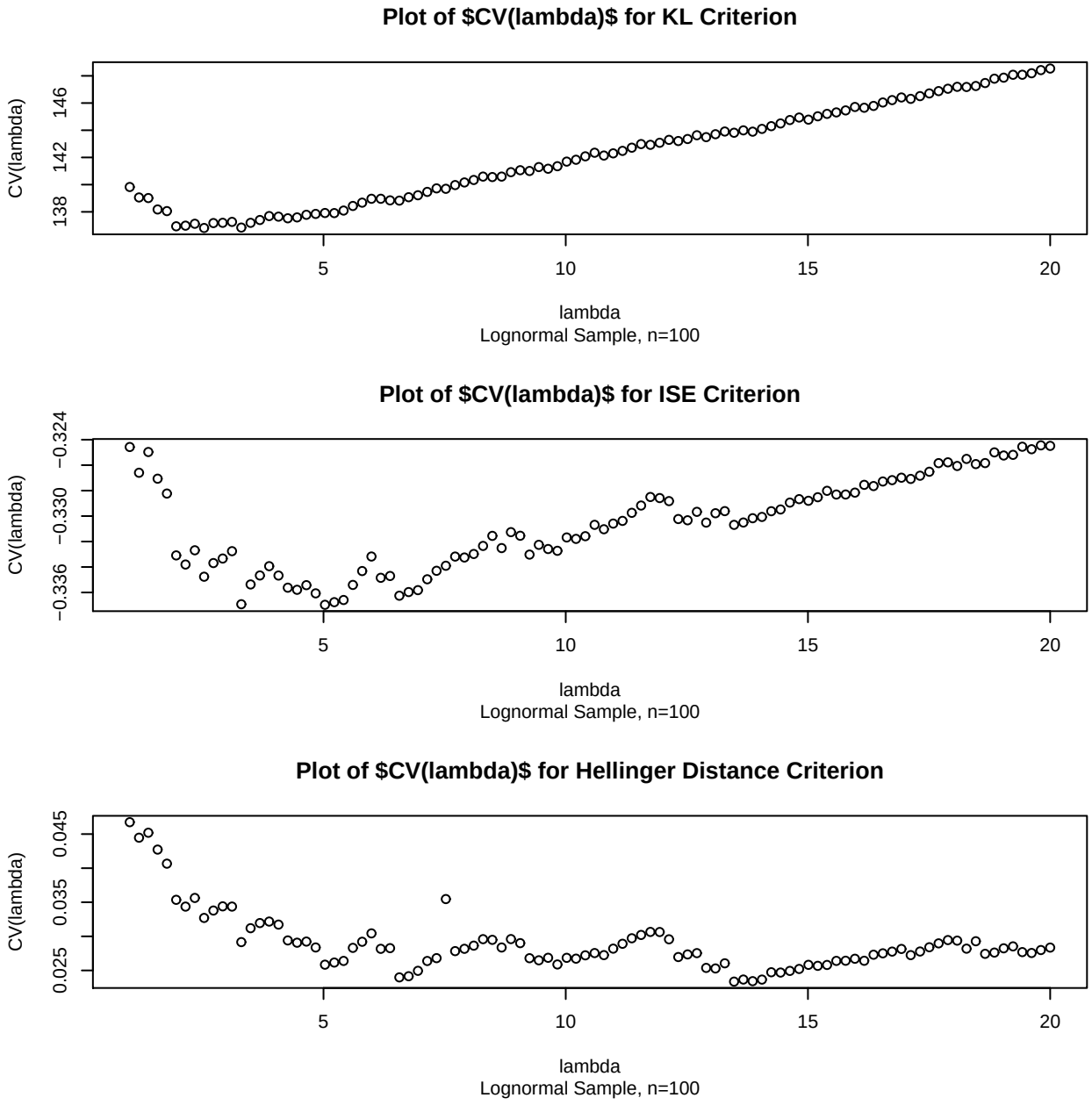

Figure 2: CV(λ) Plots for a Lognormal Sample, Sample Size = 100

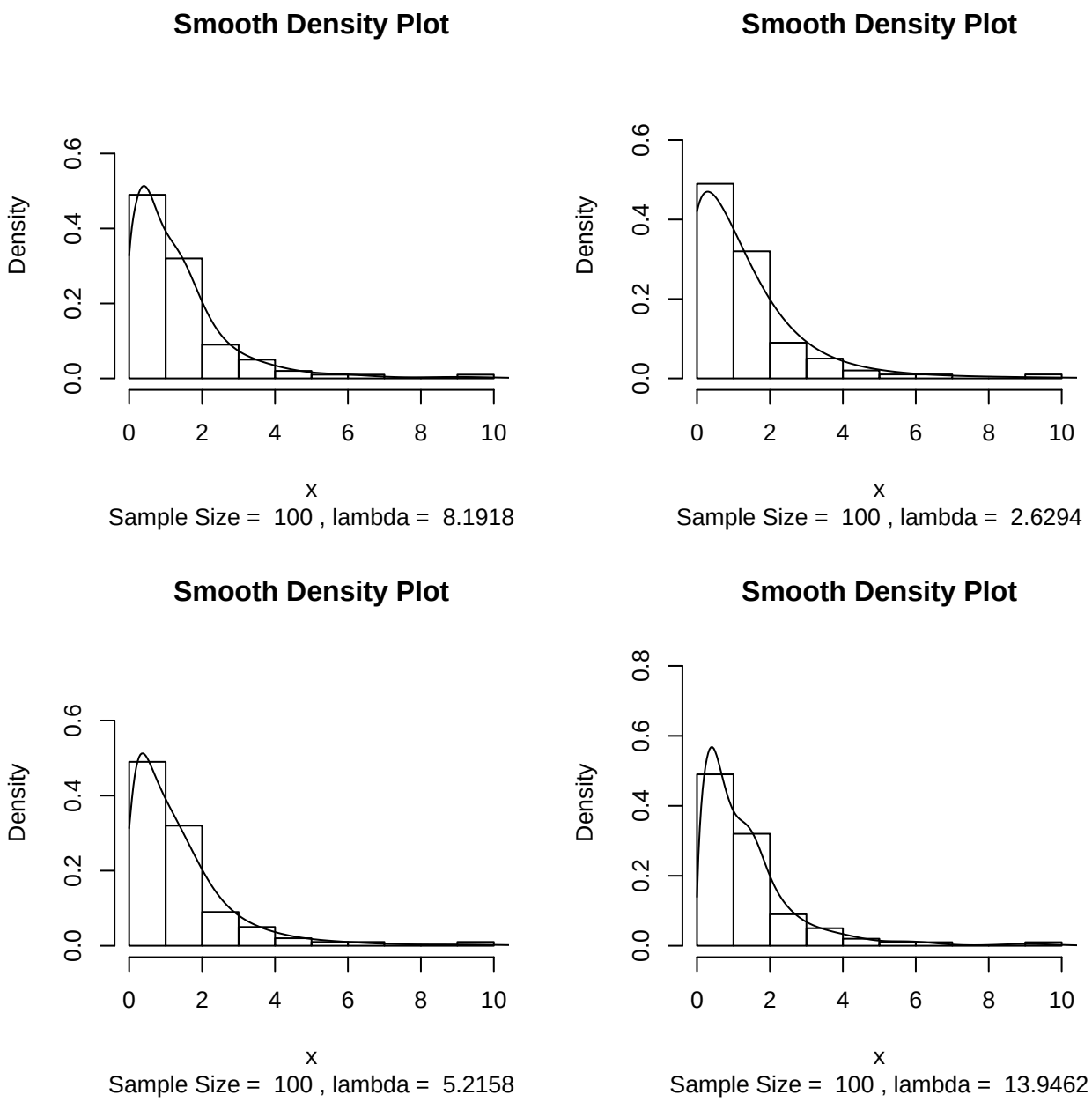

Figure 3: Smooth Density Plot for a Lognormal Sample, Sample Size = 100

3.2 Simulation for Some Standard Distributions

We have simulated from the following densities:

(1). Exponential Distribution

$$f(x) = \exp(-x)I(x > 0)$$

(2). Lognormal Distribution

$$f(x) = \frac{1}{x\sqrt{2\pi})\exp\{-\frac{1}{2}(\log x)^2}I(x > 0)$$

(3). Gamma(α) Distribution

$$f(x) = \frac{1}{\Gamma(\alpha)}\exp(-x)x^{\alpha-1}I(x > 0)$$

(4). Weibull(α) Distribution

$$f(x) = \alpha x^{\alpha-1}\exp(-x^{\alpha})I(x > 0)$$

(5). Mixtures of two Exponential Distributions

$$f(x) = [\pi\frac{1}{\theta_1}\exp(-x/\theta_1) + (1-\pi)\frac{1}{\theta_1}\exp(-x/\theta_1)]I(x > 0),$$

where we choose $\theta_1 \geq \theta_2$ and $\pi \neq 0.5$. In the simulations, I have fixed $\theta_1 = 1$. Note that we have generally not incorporated the scales in these distributions, because of the following invariance property. Denote by $\tilde{f}_{nX}(, lambda_n)$ as the smooth density based on X data using parameter λ_n. Suppose that X goes through a scale transformation $Y = X/c$ where c is a positive constant. Then it can be easily seen that

$$\begin{aligned}
\tilde{f}_{nY}(y;?) &= c\tilde{f}_{nX}(cy;\lambda_n) \\
&= c\lambda_n\sum_{j=0}^{N} p_j(\lambda_n cy)w_j(\lambda_n,\mathcal{D}) \\
&= c\lambda_n\sum_{j=0}^{N} p_j(\lambda_n cy)[G_n(\frac{j+1}{c\lambda_n}) - G_n(\frac{j}{c\lambda_n})] \\
&= \tilde{f}_{nY}(y;\lambda_n^*),
\end{aligned}$$

where $\lambda_n^* = c\lambda_n$, here G_n denotes the *edf* of the transformed data $X_1/c, ..., X_n/c$.

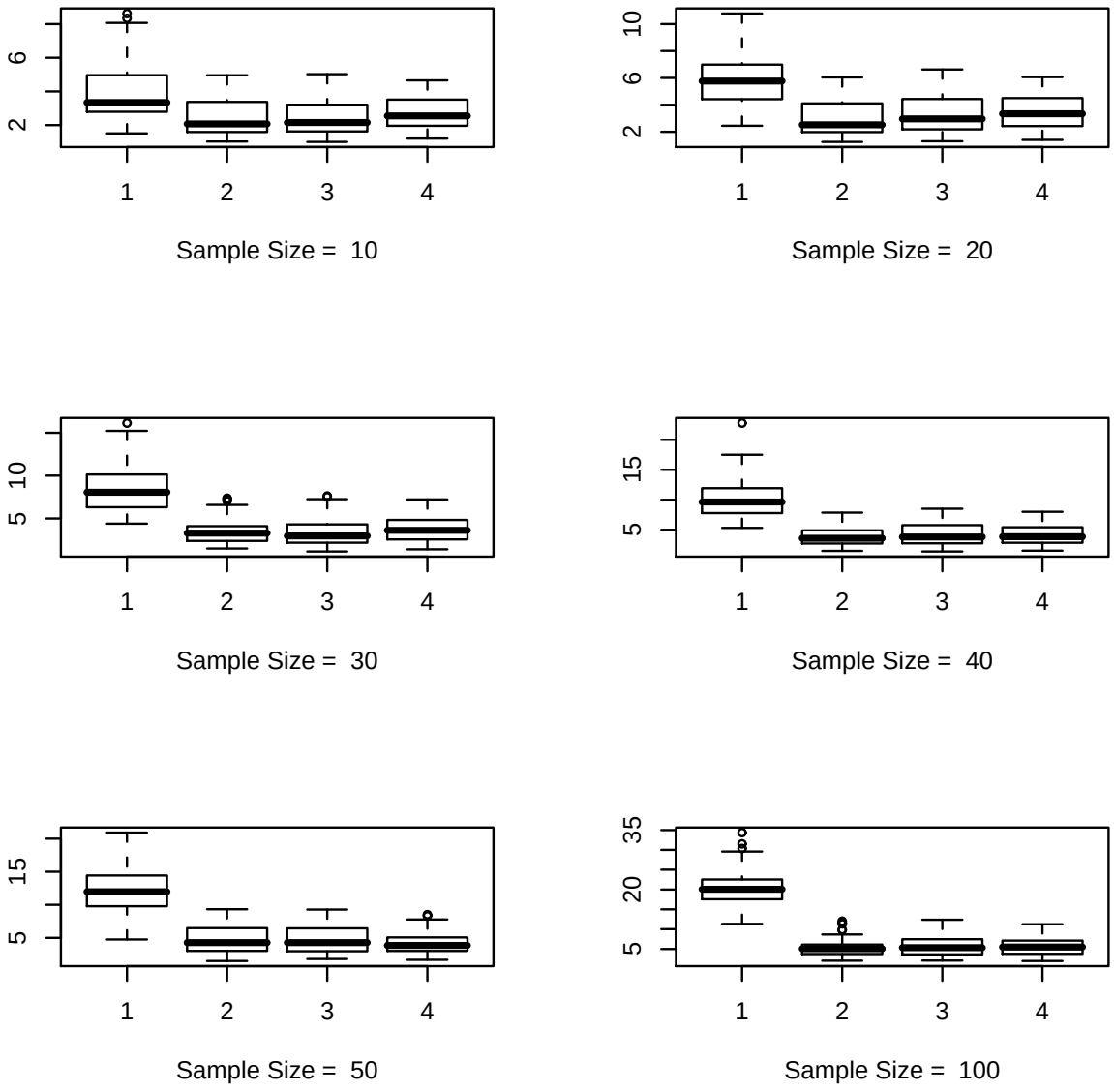

Figure 4: Box Plot for λ_n for 100 Exponential Samples 1: For Chaubey-Sen Choice, 2: KL Cross-Validation, 3: ISE Cross-Validations 4: Optimum Hellinger Distance

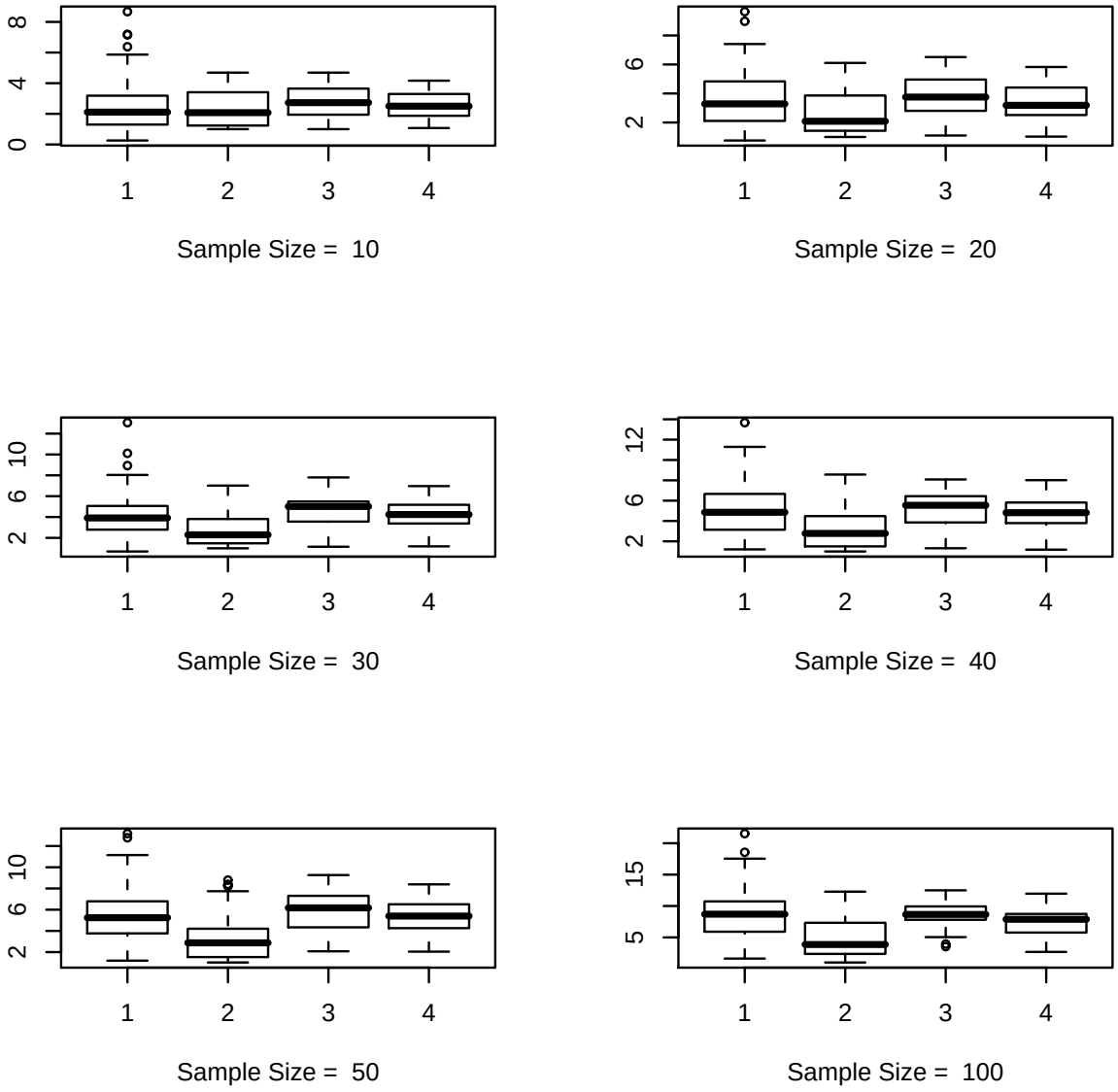

Figure 5: Box Plot for λ_n for 100 Lognormal Samples 1: For Chaubey-Sen Choice, 2: KL Cross-Validation, 3: ISE Cross-Validations 4: Optimum Hellinger Distance

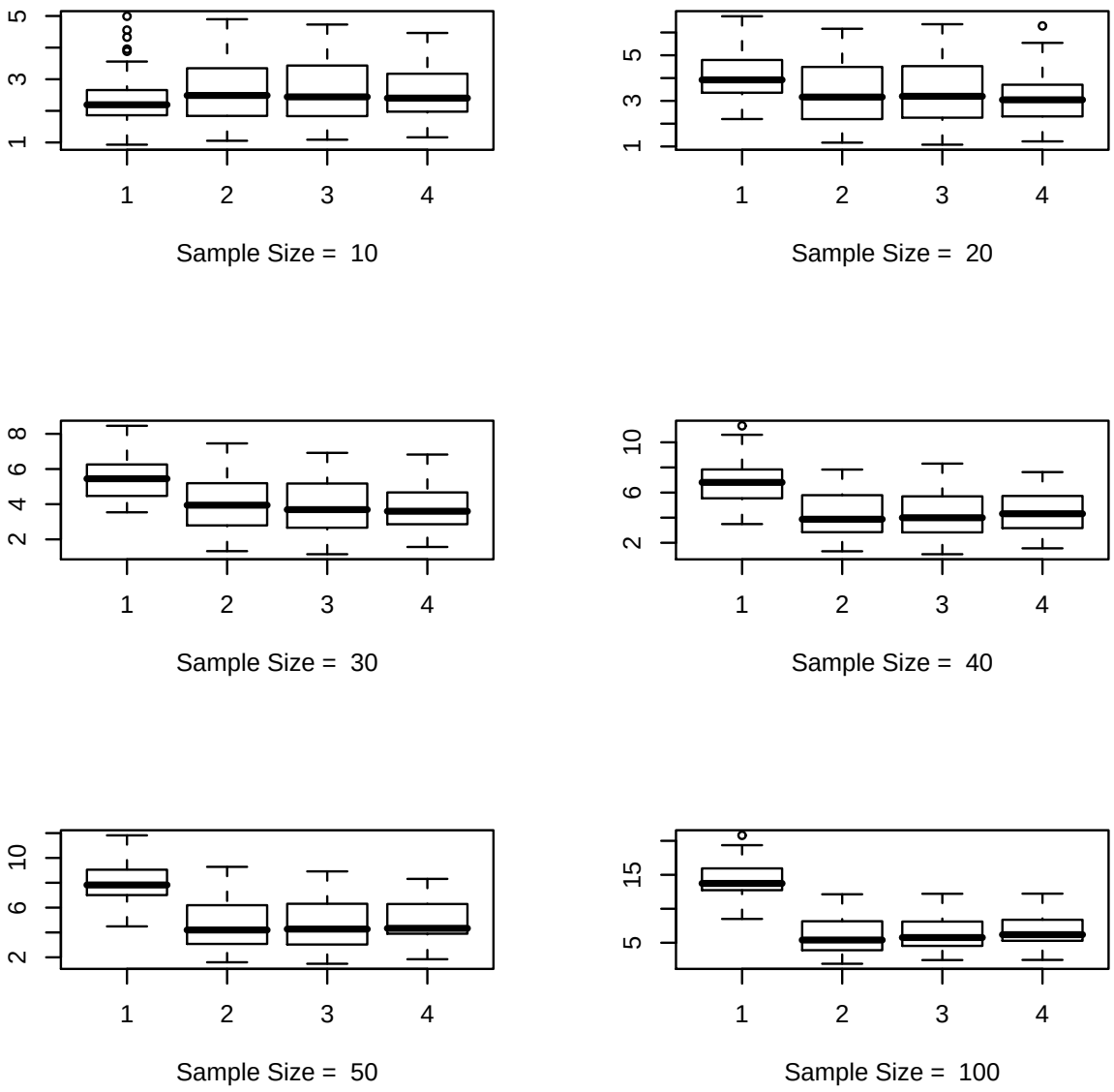

Figure 6: Box Plot for λ_n for 100 Gamma Samples, $\alpha = 2$, 1: For Chaubey-Sen Choice, 2: KL Cross-Validation, 3: ISE Cross-Validations 4: Optimum Hellinger Distance

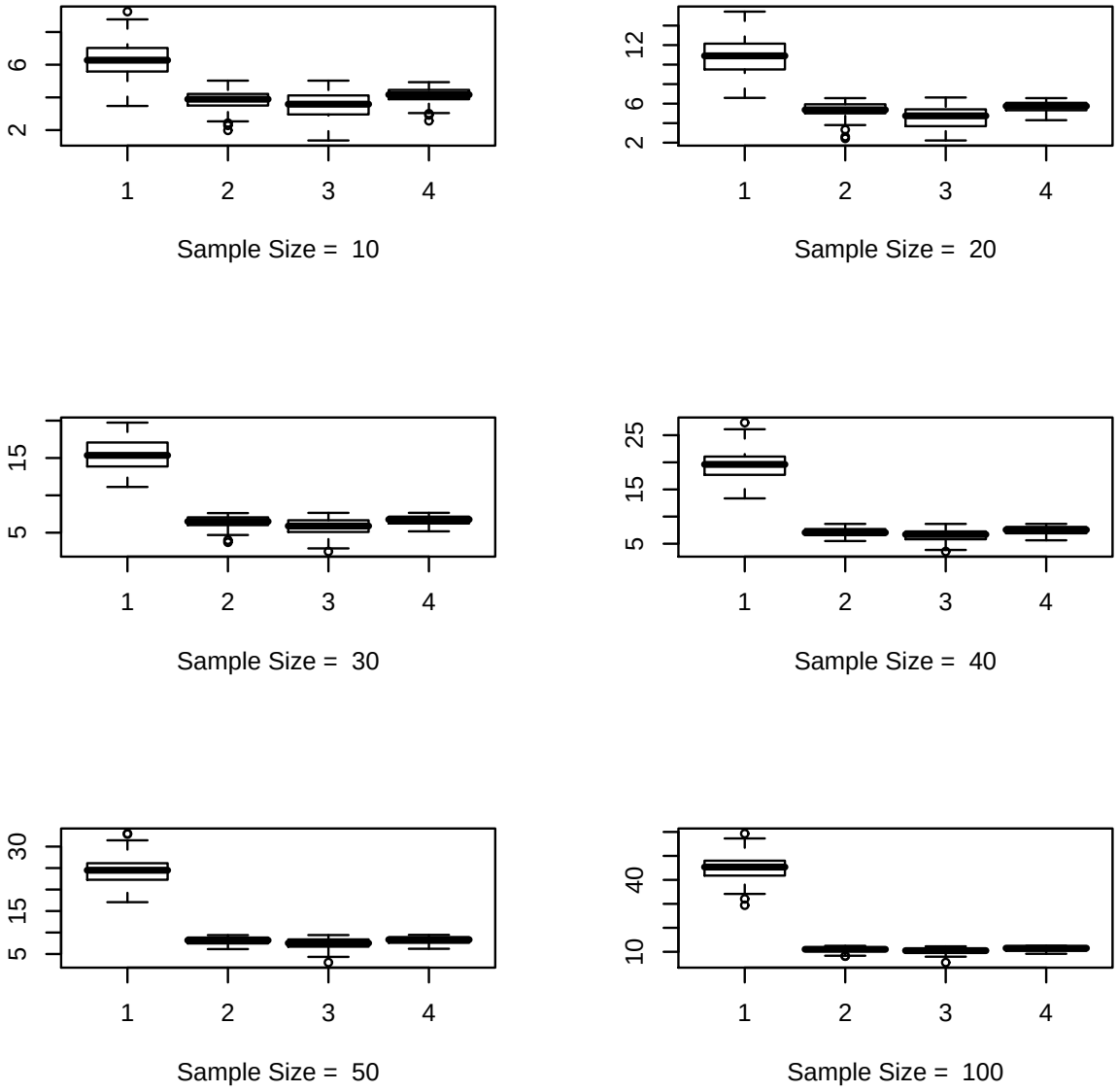

Figure 7: Box Plot for λ_n for 100 Weibull Samples, $\alpha = 2, 1$: For Chaubey-Sen Choice, 2: KL Cross-Validation, 3: ISE Cross-Validations 4: Optimum Hellinger Distance

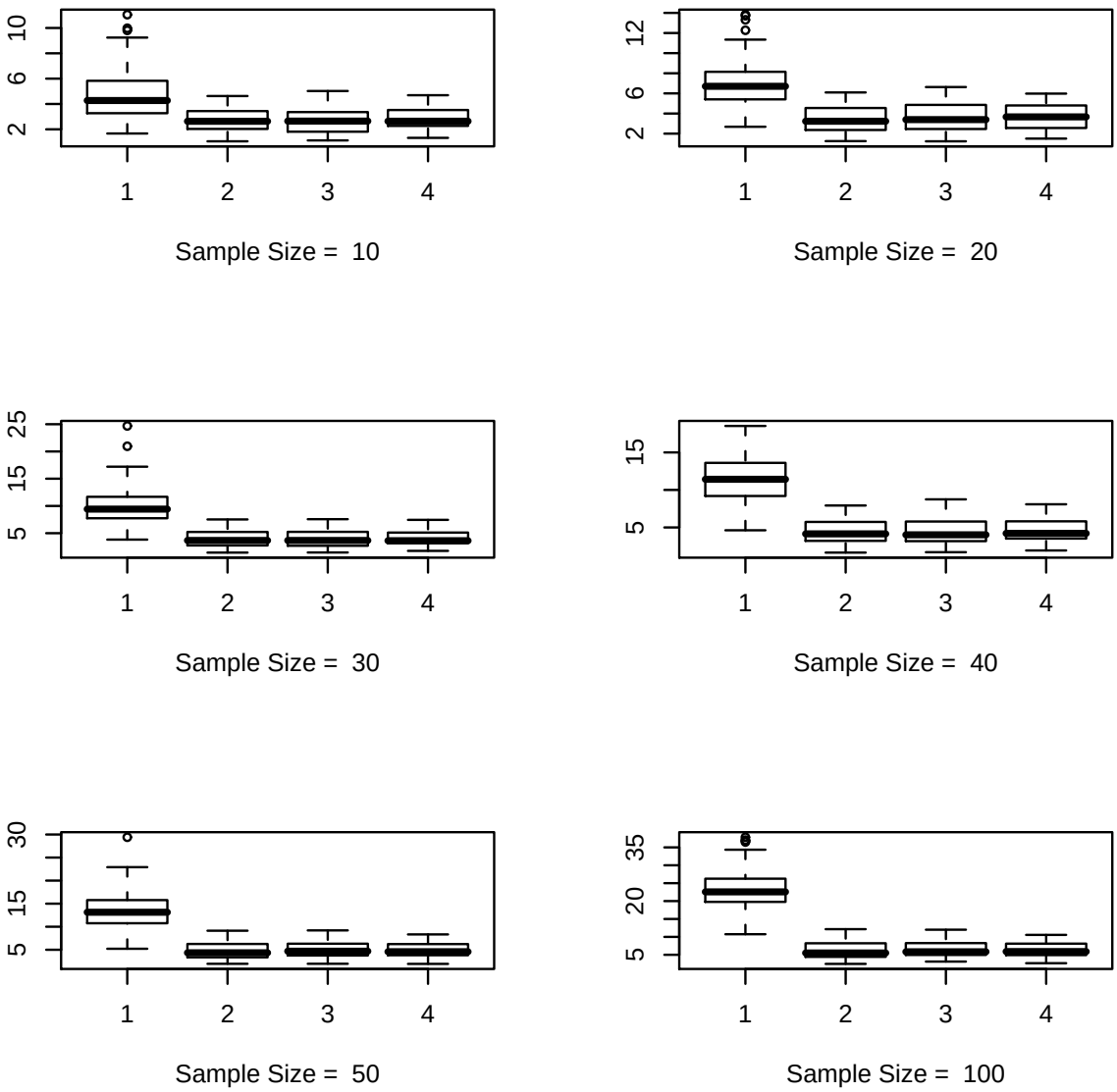

Figure 8: Box Plot for λ_n for 100 Exponential Mixture Samples, $\theta_1 = 2, \theta_2 = 1, \Pi = .4$, 1: For Chaubey-Sen Choice, 2: KL Cross-Validation, 3: ISE Cross-Validations 4: Optimum Hellinger Distance

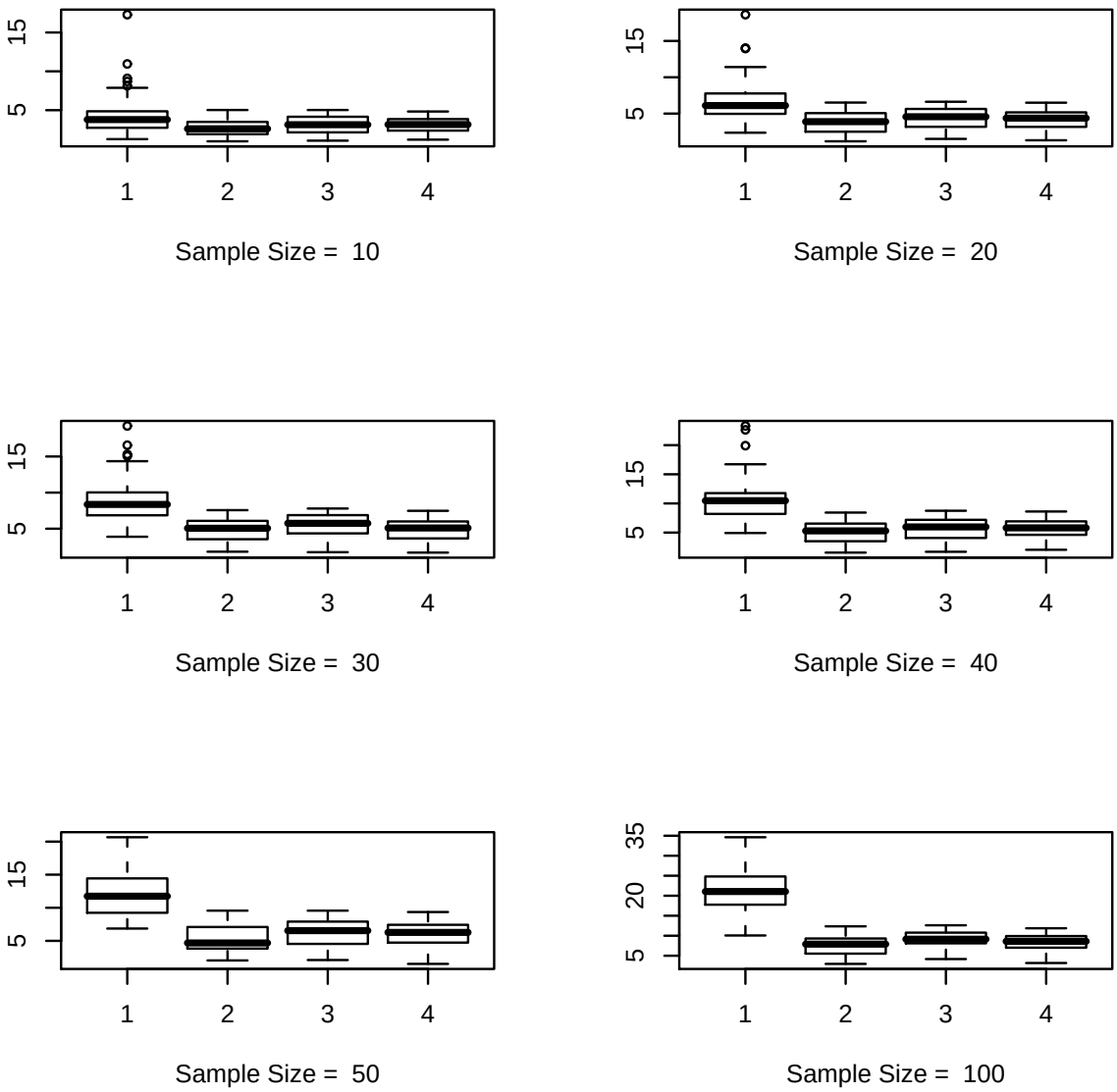

Figure 9: Box Plot for λ_n for 100 Exponential Mixture Samples, $\theta_1 = 10, \theta_2 = 1, \Pi = .2$, 1: For Chaubey-Sen Choice, 2: KL Cross-Validation, 3: ISE Cross-Validations 4: Optimum Hellinger Distance

4 Conclusion

Denote by λ_{1O} the value which minimizes the Kullback-Liebler divergence

$$KL(\lambda_n) = \mathbb{E} \int \log \frac{f(x)}{\tilde{f}_n(x)} dF(x),$$

λ_{2O} the minimizer of

$$MISE(\lambda_n) = \mathbb{E} \int (\tilde{f}_n(x) - f(x))^2 dx$$

and λ_{3O} the minimizer of the expected Hellinger distance,

$$h(\lambda_n) = \mathbb{E} \int (\sqrt{\tilde{f}_n(x)} - \sqrt{f(x)})^2 dx.$$

We have suppressed the index n, in λ_{iO} to differentiate it from the stochastic data dependent choice λ_{in}.

It is seen that

1. Chaubey-Sen choice usually produces large values of the smoothing parameters, especially, for large samples. Because of the invariance property of the estimator, choice of the scale of the data does not affect the optimum value.

2. Chaubey-Sen choice is much more variable even in the cases on an average it is close to the true optimum.

3. The two cross-validation criteria generally produce similar results, especially for larger samples and they converge to the true optimum under the known density.

4. We conjecture that suppose λ_{iO} denotes the true value of λ_n which minimizes criterion $i, i = 1, 2, 3$, and λ_{in} is the minima based on the data, then

$$(i) \lim_{n\to\infty} \frac{\lambda_{in}}{\lambda_{iO}} = 1 \quad \text{a.s.}$$

$$(ii) \lim_{n\to\infty} \frac{\lambda_{1O}}{\lambda_{HO}} = \lim_{n\to\infty} \frac{\lambda_{2O}}{\lambda_{HO}} = 1 \quad \text{a.s.},$$

where λ_{HO} is the true minimizer of the expected Hellinger distance between $\tilde{f}_n$ and f.

List of Recent Technical Reports

77. Alexander Melnikov and Victoria Skornyakova, *Efficient Hedging Methodology Applied to Equity–Linked Life Insurance*, February 2005

78. Qihe Tang, *The Finite Time Ruin Probability of the Compound Poisson Model with Constant Interest Force*, June 2005

79. Marc J. Goovaerts, Rob Kaas, Roger J.A. Laeven, Qihe Tang and Raluca Vernic, *The Tail Probability of Discounted Sums of Pareto–Like Losses in Insurance*, August 2005

80. Yogendra P. Chaubey and Haipeng Xu, *Smooth Estimation of Survival Functions under Mean Residual Life Ordering*, August 2005

81. Xiaowen Zhou, *Stepping–Stone Model with Circular Brownian Migration*, August 2005

82. José Garrido and Manuel Morales, *On the Expected Discounted Penalty Function for Lévy Risk Processes*, November 2005

83. Ze–Chun Hu, Zhi–Ming Ma and Wei Sun, *Extensions of Lévy–Khintchine Formula and Beurling–Deny Formula in Semi–Dirichlet Forms Setting*, February 2006

84. Ze–Chun Hu, Zhi–Ming Ma and Wei Sun, *Formulae of Beurling–Deny and Lejan for Non-Symmetric Dirichlet Forms*, February 2006

85. Ze–Chun Hu and Wei Sun, *A Note on Exponential Stability of the Non-Linear Filter for Denumerable Markov Chains*, February 2006

86. H. Brito–Santana, R. Rodríguez–Ramos, R. Guinovart–Díaz, J. Bravo–Castillero and F.J. Sabina, *Variational Bounds for Multiphase Transversely Isotropic Composites*, August 2006

87. José Garrido and Jun Zhou, *Credibility Theory for Generalized Linear and Mixed Models*, December 2006

88. Daniel Dufresne, José Garrido and Manuel Morales, *Fourier Inversion Formulas in Option Pricing and Insurance*, December 2006

89. Xiaowen Zhou, *A Superprocess Involving Both Branching and Coalescing*, December 2006

90. Yogendra P. Chaubey, Arusharka Sen and Pranab K. Sen, *A New Smooth Density Estimator for Non–Negative Random Variables*, January 2007

91. Md. Sharif Mozumder and José Garrido, *On the Relation between the Lévy Measure and the Jump Function of a Lévy Process*, October 2007

92. Arusharka Sen and Winfried Stute, *A Bi-Variate Kaplan-Meier Estimator via an Integral Equation*, October 2007

93. C. Sangüesa, *Uniform Error Bounds in Continuous Approximations of Nonnegative Random Variables Using Laplace Transforms*, January 2008

94. Yogendra P. Chaubey, Naâmane Laib and Arusharka Sen, *A Smooth Estimator of Regression Function for Non-negative Dependent Random Variables*, March 2008

95. Alejandro Balbás, Beatriz Balbás and Antonio Heras, *Optimal Reinsurance with General Risk Functions*, March 2008

96. Alejandro Balbás, Raquel Balbás and José Garrido, *Extending Pricing Rules with General Risk Functions*, April 2008

97. Yogendra P. Chaubey and Pranab K. Sen, *On the Selection of the Smoothing Parameter in Poisson Smoothing of Histogram Estimator: Computational Aspects*, December 2008

Copies of technical reports can be requested from:

Dr. Wei Sun
Department of Mathematics and Statistics
Concordia University
1455 de Maisonneuve Blvd. West,
Montreal, QC, H3G 1M8 CANADA